Assignment2 - Step 6

Phoebe Zhou

September 2025

Advanced Evaluation with RAGAs

I compared two Retrieval-Augmented Generation (RAG) pipelines: a naive version, which just retrieves documents and feeds them to the model, and an enhanced version, which adds query rewriting and reranking to improve retrieval quality. To see how much the enhancements helped, I used RAGAs to measure four key metrics: faithfulness, answer relevancy, context precision, and context recall.

Metric	Naive	Enhanced	Change
Faithfulness	0.8562	0.9385	+0.0823
Answer Relevancy	0.7388	0.7746	+0.0358
Context Precision	0.6931	0.8542	+0.1610
Context Recall	0.6562	0.8047	+0.1484

1. Faithfulness: The enhanced version did a better job of sticking to the facts in the documents. This is huge for short answers — nobody wants a confident "yes" when the real answer is "no."
2. Answer relevancy: The gain here was smaller, but still positive. Since most questions are very clear, both pipelines gave relevant answers most of the time.
3. Context precision: This was the biggest jump. The enhanced pipeline retrieved cleaner, more on-target passages instead of grabbing extra noise. That made the answers sharper and more reliable.
4. Context recall: Also improved a lot. The enhanced pipeline pulled in more of the relevant evidence, which matters when multiple documents mention the same fact.

Takeaways

The naive pipeline wasn't terrible as it handled straightforward questions decently. But the enhanced pipeline consistently did better, especially in precision and recall of context, which had the biggest impact on accuracy. For short, fact-style questions, these improvements make the answers feel more trustworthy and less error-prone.

In short: the enhanced RAG is clearly worth it. It gives better evidence, sticks closer to the facts, and delivers more accurate answers overall.